UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Modelo Dinâmico em Seguros para Negócio novo e continuado

Ana Madalena Coimbra Carmelo

Mestrado em Matemática Aplicada à Economia e Gestão

Versão Pública

Trabalho de Projeto orientado por:
Prof. Dr. João José Ferreira Gomes

2020

“The answer will come to him who tries to look at his life through heaven's eyes.”

Stephen Schwartz

Agradecimentos

Em primeiro lugar gostaria de agradecer ao meu supervisor externo, da parte da Ageas Seguros, João Pedro Oliveira, por ter confiado em mim e ter incentivado o desenvolvimento do meu potencial do primeiro ao último dia da realização deste estágio.

Ao meu supervisor interno, o Professor João José Ferreira Gomes, agradeço o apoio, a disponibilidade e a forma com que contribuiu para o agilizar deste projeto, colocando tantas vezes o melhor para o meu futuro como prioridade. Por ter aceite ser meu orientador com tal entusiasmo, um enorme obrigada.

Finalmente, um dia a minha avó materna, extremamente religiosa, perguntou qual a origem do meu sossego nos momentos de maior desespero. Ao que eu lhe respondi: Vocês. Vocês são a minha fé.
Por isso, dedico este trabalho ao núcleo de pessoas, que para mim, representa uma mão invisível pronta a amparar a queda bem como a empurrar-me na altura de voar. À minha mãe e Acácio, ao meu pai, à minha irmã, avós, melhores amigas e ao Teddy.

Resumo

A atividade seguradora em Portugal está em crescimento. Este setor é fundamental para a estabilidade da economia nacional, uma vez que abrange diversas áreas da vida dos seres humanos. Sejam essas a saúde, os bens, o financiamento ou a poupança. Associado ao crescente número de clientes está a competição entre entidades seguradoras para obter a liderança do mercado. Tendo esse objetivo, uma das principais funções das seguradoras é estudar as caraterísticas dos seus potenciais clientes.

A variação do prémio é a variável que a seguradora pode alterar. Ao estudar a reação dos vários tipos de clientes face às alterações na tarifa, esta poderá encontrar o valor do prémio mais favorável à existência de uma harmonia entre a satisfação do cliente e o lucro da empresa.

A concretização deste estágio insere-se num ramo particular do mercado segurador e teve como objetivo a análise da dinâmica de dois tipos de contratos, os contratos novos e os contratos continuados. Isto é, os contratos de clientes novos na empresa e os contratos de clientes que já existiam na empresa e pretendem renovar o contrato no fim da sua vigência. Desse modo, estudou-se a elasticidade dos clientes através da análise da taxa de conversão nos contratos novos e a probabilidade de renovação, através da construção de um modelo de renovação, nos contratos continuados. Contudo, por motivos de confidencialidade, não foi possível apresentar nesta versão do relatório, os resultados derivados da análise dos contratos novos.

Na construção do modelo de renovação utilizaram-se as árvores de decisão e ainda a regressão Logística, para modelar a variável de resposta. A variável de resposta é uma variável binária que toma os valores 0 e 1, consoante o cliente renove ou não o contrato, respetivamente. No fim, compararam-se os vários modelos obtidos, e selecionou-se o mais adequado, de acordo com a análise dos índices relevantes, obtendo-se assim a panóplia de variáveis que influenciam a renovação. Essa descoberta permitiu a criação de alguns perfis de renovação.

PALAVRAS-CHAVE: Seguro Empresas, Taxa de Conversão, Elasticidade, Modelo de Regressão Logística.

Abstract

Insurance activity is constantly growing in Portugal. This sector is fundamental for the stability of the national economy, since it covers several areas of human beings' lives. Whether those are health, assets, financing or savings areas. Associated with the growing number of customers is the competition between insurance companies to obtain the market leadership. With this objective in mind, one of the main functions of the insurance companies is to study the characteristics of their potential customers.

Since the premium variation is the only variable that the insurance company is able to control, when studying the reaction of the various types of customers in the face of changes in the tariff, the company can find the premium value which results in the best balance between customer satisfaction and the company's profit.

This internship is integrated in a specific branch and aimed to analyze the dynamics of two types of contracts, new contracts and continuing contracts. Those are, contracts for new customers in the company and contracts for customers that already existed in the company and intend to renew the contract at the end of its term. The elasticity of customers was studied through the analysis of the conversion rate, in new contracts, and the likelihood of renewal, through the construction of a renewal model, in continuing contracts. However, due to the data confidentiality, it was not possible to include the results of the new contracts analysis in this report version.

In the construction of the renewal model, decision trees and Logistic regression were used to model the response variable. The response variable was a binary variable that takes the values 0 and 1 depending on whether the customer renews the contract or not, respectively. In the end, the various models obtained were compared, and the most appropriate was selected, according to the analysis of the relevant performance indicators, thus obtaining the panoply of variables that influence the renewal. This discovery allowed the creation of some renewal profiles.

Keywords: Multirisk Insurance, Conversion Rate, Elasticity, Logistic Regression Model

Índice

Lista Gráficos

Lista de Tabelas

Lista de Figuras

Glossário e Abreviaturas[1]

APÓLICE: conjunto de Condições identificado na cláusula anterior e na qual é formalizado o contrato de seguro celebrado;
AGENTE: Mediador que pode exercer atividade para uma ou mais companhias de seguros, onde angaria clientes, apresenta soluções e propostas, que posteriormente são traduzidas em contratos de seguro.
ASF: Autoridade de Supervisão de Seguros e Fundos de Pensões, entidade responsável pela regulação e supervisão das seguradoras, resseguradoras, fundos de pensões e mediação de seguros.
SEGURADOR: a entidade legalmente autorizada para a exploração do seguro obrigatório de incêndio, que subscreve o presente contrato;
TOMADOR DO SEGURO: a pessoa ou entidade que contrata com o Segurador, sendo responsável pelo pagamento do prémio;
SEGURADO: a pessoa ou entidade titular do interesse seguro;
BENEFICIÁRIO: a pessoa ou entidade a favor de quem reverte a prestação do Segurador por efeito da cobertura prevista no contrato;
INCÊNDIO: a combustão acidental, com desenvolvimento de chamas, estranha a uma fonte normal de fogo, ainda que nesta possa ter origem, e que se pode propagar pelos seus próprios meios;
AÇÃO MECÂNICA DE QUEDA DE RAIO: a descarga atmosférica ocorrida entre a nuvem e o solo, consistindo em um ou mais impulsos de corrente que confere ao fenómeno uma luminosidade característica (raio) e que provoque deformações mecânicas e permanentes nos bens seguros;
EXPLOSÃO: a ação súbita e violenta da pressão ou depressão de gás ou de vapor;
SINISTRO: a verificação, total ou parcial, do evento que desencadeia o acionamento da cobertura do risco prevista no contrato;
PRÉMIO COMERCIAL: Custo técnico das coberturas do contrato, este prémio não considera os custos relacionados com a emissão do contrato, de aquisição, de gestão ou de cobrança.
PRÉMIO TOTAL: O prémio é a contrapartida da cobertura acordada e inclui tudo o que seja contratualmente devido pelo tomador do seguro, nomeadamente os custos da cobertura do risco, os custos de aquisição, de gestão e de cobrança, custos de fracionamento, custo de Apólice/ata adicional, carta verde e dos encargos fiscais e parafiscais a suportar pelo Tomador do seguro
FRANQUIA: valor da regularização do sinistro nos termos do contrato de seguro que não fica a cargo do Segurador.
ESTABELECIMENTO SEGURO: todo aquele que, como tal, for designado e identificado na Apólice;
ACIDENTE NO ESTABELECIMENTO SEGURO: o acontecimento, fortuito, súbito e imprevisível, violento ou não, ocorrido no Estabelecimento Seguro devido a causa exterior e estranha à vontade das Pessoas Seguras, em consequência dos riscos cobertos.

[1] De acordo com a página online da empresa.

Introdução

O presente relatório insere-se no âmbito de um estágio concretizado na Ageas Seguros, mais concretamente, na equipa de Pricing and Business Analytics, na direção Técnica e Oferta Não Vida. Numa seguradora, a direção Técnica e Oferta é a responsável por criar, desenvolver e implementar estratégias para ajustar a oferta e o preço dos produtos que a seguradora oferece aos clientes, de modo a ganhar participação de mercado e alcançar metas de receitas. A equipa de Pricing & Business Analytics tem a função extremamente analítica de analisar dados de várias fontes, desenvolver modelos complexos dos preços e ainda gerir o portfólio.

Existem dois momentos decisivos numa seguradora. O momento de aquisição dos riscos e assinatura do contrato, isto é, a conversão da proposta, e o momento de renovação desse contrato. Neste projeto abordar-se-á contratos nessas duas situações.

Examinando dados anteriores, é possível prever quais as caraterísticas que definem o cliente ideal e assim prevenir grandes prejuízos. Além disso, ao identificar o perfil dos clientes que mais renovam ou não, as seguradoras são capazes de tomar as medidas necessárias para reter os clientes com menor risco, aumentando assim o seu lucro.

A ferramenta para a construção das bases de dados do projeto foi o software SAS Enterprise Guide. Para a construção do modelo de renovação utilizou-se o SAS Enterprise Miner. Neste último usou-se a regressão logística, para modelar a variável resposta.

Neste documento, no Capítulo I é introduzido o conceito de atividade seguradora, bem como uma introdução à seguradora na qual este projeto foi realizado e o ramo em que se insere. O Capítulo II, diz respeito ao enquadramento teórico que fundamentou o enquadramento técnico do projeto, mais concretamente, uma abordagem da Regressão Logística. O Capítulo III descreve a construção do modelo de renovação, as variáveis utilizadas, a correlação entre variáveis, a avaliação do modelo final e alguns perfis de renovação. Por fim, no capítulo IV apresenta-se o desfecho do projeto.

CAPÍTULO I - Atividade Seguradora

História da Atividade Seguradora em Portugal[2]

Como dita a história, o aparecimento do conceito de seguro surgiu pela primeira vez em Portugal no ano de 1293. Nessa altura governava o Rei D.Dinis, o responsável por estabelecer a primeira forma de seguro. O primeiro seguro a existir estava relacionado com as descobertas marítimas. De modo a garantir a salvaguarda das suas embarcações, os mercadores celebravam um contrato entre si consoante o pagamento de uma quantia – o Prémio – para fazer face aos custos em caso de perda do navio ou da mercadoria.
Em 1370 foram fixadas as primeiras leis sobre seguros pelo rei D.Fernando I e em 1380 foi criada a Companhia das Naus pelo mesmo. A companhia das Naus funcionava como uma companhia de seguros, dando aos proprietários de navios uma salvaguarda em caso de sinistro. O surgimento desta companhia teve uma elevada importância no desenvolvimento da área seguradora uma vez que impulsionou o desenvolvimento da marinha Portuguesa. Com o desenvolvimento da marinha tomou-se perceção da importância da existência dos seguros e assim sendo em 1383 estabeleceu-se a primeira lei nacional sobre Seguros.
À data de 1578 foi criada a primeira entidade reguladora de seguros, o corretor de seguros. Esta entidade é responsável por validar os seguros, assegurando os direitos do segurador.
A primeira companhia Portuguesa de seguros – Companhia permanente de Seguros – foi criada em 1791, levando ao aparecimento de cada vez mais companhias de seguros nos anos seguintes.
Por fim, em 1982, surgiu a Associação Portuguesa de Seguradores - APS.

Grupo AGEAS

AGEAS[3]

A AGEAS é um grupo segurador internacional, com sede em Bruxelas, Bélgica. A empresa surgiu na época da crise financeira de 2008 e 2009 a partir da venda do grupo de serviços financeiros Fortis. Neste momento está presente em 15 países da Europa e da Ásia, onde a empresa propõe soluções de seguros de Vida e Não Vida a milhões de Clientes particulares e empresas. É uma das maiores seguradoras europeias e é a seguradora líder no seu país de origem, Bélgica.

AGEAS Portugal

O grupo Ageas entrou no mercado segurador Português em 2005, através de uma *joint venture* com o Banco Millennium BCP. Neste momento é considerada uma das empresas líderes no ranking segurador português. Desde 2005 que opera através de marcas conhecidas como a Ocidental e a Médis. Em abril de 2016 o grupo AGEAS adquiriu a AXA Portugal (atualmente denominada como Ageas Seguros) e a Seguro Direto. A empresa reforçou assim o seu investimento no mercado Português, criando os canais de distribuição com uma rede de mediadores Profissionais e um canal de veda direta. A Ageas Seguros opera nos segmentos de Vida e Não Vida. O seu objetivo passa por contribuir para o desenvolvimento do país e da sociedade e servir os seus clientes com um vasto leque de ramos, para que este possa antecipar e proteger-se contra riscos imprevisíveis.

[2] De acordo com a página online historiadoseguro.com.
[3] De acordo com a página online da empresa.

Atividade Seguradora

Contrato de Seguro[4]

O contrato de seguro é um acordo através do qual o segurador se compromete a cobrir determinados riscos, assumindo o pagamento das indemnizações ou do capital seguro, em caso de ocorrência de sinistro, a realização total ou parcial do risco, nos termos acordados. Por sua vez, a pessoa ou entidade segurada, isto é, o tomador do seguro, fica obrigado a pagar ao segurador o prémio correspondente ao custo do seguro. A prestação acordada no contrato pode ser paga ao segurado, a um terceiro designado pelo tomador do seguro, o beneficiário, ou ainda, a uma terceira pessoa ou entidade que tenha sofrido prejuízos que o segurado deva indemnizar – o terceiro lesado. Quanto à obrigatoriedade os seguros podem ser obrigatórios ou facultativos consoante a exigência por lei.

Contratos que devido à sua grande dimensão, ou à sua natureza apresentam um nível de risco elevado, necessitam da intervenção de várias seguradoras. Nesses casos, existem os designados contratos de co-seguro. Esses são contratos celebrados entre várias seguradoras, em que cada co-seguradora tem uma participação da quantia segurada. A co-seguradora líder, é a seguradora com maior responsabilidade no contrato.

Tipos de Seguros

A atividade seguradora divide-se em dois grandes ramos: Seguros Vida e Seguros Não Vida.

A divisão dos Seguros de Vida é:

- Seguro de Vida – Seguro realizado sobre a vida de uma ou mais pessoas e cobre principalmente, o risco de morte ou o risco de sobrevivência;
- Seguro de Nupcialidade – Pagamento de um prémio ou de uma renda caso a pessoa segura se case;
- Seguro de Natalidade – Pagamento de um prémio ou de uma renda em caso de nascimento de filhos;
- Seguro de Fundos de Investimento Coletivo – Seguro em que as importâncias seguras são determinadas em função de um "valor de referência constituído por uma "unidade de conta" ou pela combinação de várias unidades de conta;

A divisão dos Seguros de Não Vida é:

- Automóvel – Seguro que se divide em duas componentes, Responsabilidade Civil e Danos Próprios.
- Acidentes de Trabalho – Visa a proteção do indivíduo contra qualquer acidente que ocorra no seu local de trabalho ou no trajeto para o trabalho;
- Acidentes Pessoais – Seguro que prevê a proteção do indivíduo contra qualquer acidente que ocorra no decurso do seu dia e que esteja fora do âmbito dos seguros obrigatórios;
- Saúde - Seguro que assume riscos relacionados com a prestação de cuidados de saúde, dependendo das coberturas visadas nas condições do contrato;

[4] De acordo com a página online da Autoridade de Supervisão de Seguros e Fundos de Pensões.

- Multirriscos – Seguro que cobre os riscos relacionados com a habitação ou empresas. Consoante as coberturas que o cliente subscrever, pode garantir, que caso ocorra algum imprevisto, o imóvel e os bens que este contenha ficarão segurados.

Existe um conjunto de seguros obrigatórios daqueles apresentados, entre outros. Alguns deles são:

- Acidentes de trabalho;
- Automóvel – Responsabilidade civil;
- Caçador – Responsabilidade civil;
- Escolar - Acidentes pessoais;

Processo Renovações Multirriscos Empresa Ageas Seguros

O processo de renovação é um dos processos mais importantes numa companhia de seguros, uma vez que é nesse processo que a empresa determina os preços para a anuidade seguinte do contrato. No entanto, a forma de processar as renovações nem sempre é transversal a todas as companhias de seguros. Na Ageas Seguros, 45 dias antes da potencial renovação do contrato, são consideradas todas as apólices em vigor. A essa data a empresa envia ao cliente a informação do valor do prémio para a anuidade seguinte. As mesmas apólices (potenciais renovações de contrato), são analisadas 50 dias após a data de renovação e se a apólice se mantiver em vigor conclui-se que o contrato foi renovado. Por outro lado, se a apólice estiver anulada conclui-se que o contrato não foi renovado com sucesso. Dos contratos cujo pagamento se efetua com fracionamento mensal, apenas se consideram anulados os contratos em que foi paga a primeira mensalidade, mas não a segunda. Caso a anulação seja feita após os 50 dias seguintes à data de renovação, define-se que a anulação não se deve à oscilação do prémio gerada pelo processo de renovação e, portanto, não são relevantes para esta análise.

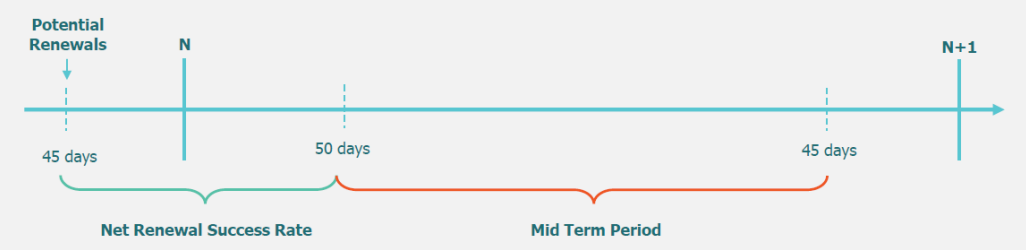

Figura 1.1 – Processo de renovação

1) *Potential Renewals* – Apólices em vigor no momento do processo de renovação, momento no qual é enviada uma carta ao cliente a informar a aproximação da data de renovação.

2) *Net Renewal Success Rate (NRSR)* – Percentagem de *Potential Renewals* que efetivamente renovaram no período de 50 dias depois do começo da anuidade. Uma apólice é considerada renovada se e só se cumpre as seguintes condições:

   - A primeira prestação é paga;
   - Não existe data de cancelamento após 50 dias do início da nova anuidade.

3) *Mid Term Cancellation Rate* – Percentagem de apólices canceladas durante o *Mid Term Period,* isto é, 50 dias após a renovação e até 45 dias antes da renovação seguinte.

CAPÍTULO II – ENQUADRAMENTO TEÓRICO

Árvores de Decisão

As árvores de decisão são um método não paramétrico usado na inferência estatística em problemas de classificação e regressão e podem ser aplicadas a dados categóricos ou numéricos. As árvores aprendem com os dados, utilizando para esse fim um conjunto de regras de decisão. Quanto mais profunda for a árvore, mais complexas são as regras de decisão e mais adequado é o modelo. Os elementos de uma árvore de decisão são os nós, os arcos e as folhas. Os nós representam as questões que se colocam sobre os dados, os arcos, o elo de ligação entre o conjunto de dados e a resposta, e por fim, as folhas, que representam os nós finais e a resposta. A partir de uma base de dados de treino, o algoritmo básico de indução das árvores de decisão divide o conjunto de dados em partições alternativas e considerando a qualidade da partição, seleciona a melhor partição. Para a partição selecionada, o processo repete-se para cada um dos elementos da partição. O algoritmo termina quando algum critério é atingido. As árvores de decisão são de fácil interpretação, percebendo-se nitidamente a razão da decisão. A escolha dos atributos mais importantes é automática e é visualmente percetível, sendo que os atributos mais relevantes aparecem sempre mais acima na árvore.

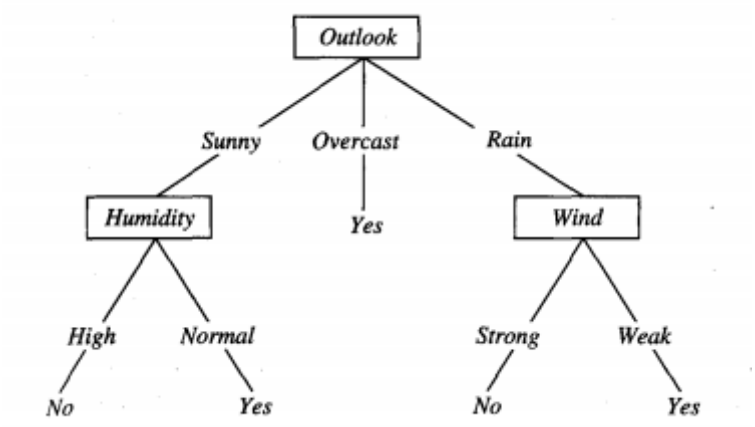

Figura 2.1 – Exemplo de uma árvore de decisão: Decisão de jogar ténis[5]

Modelos Lineares generalizados

Os Modelos Lineares Generalizados, apresentados por Nelder e Wedderburn (1972), são um conjunto de modelos de regressão, que têm como objetivo o estudo da relação entre variáveis explicativas com a variável resposta. Comummente, os modelos lineares generalizados são bastante usados na área seguradora, uma vez que as suas bases de dados, não apresentam uma distribuição normal. Na realidade, os Modelos Lineares Generalizados pressupõem que a variável resposta tem uma distribuição pertencente à família Exponencial.

Existem três etapas essenciais que se devem seguir ao tentar modelar dados através de um MLG:

- Formulação dos modelos;

[5] Mitchell, TM: 1997, Machine Learning, McGraw-Hill.

- Ajustamento dos modelos;
- Seleção e validação dos modelos;

Família Exponencial[6]

Uma variável aleatória Y diz-se ter uma distribuição exponencial se a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever da seguinte forma:

$$f(\gamma|\phi) = exp\left\{\frac{\gamma\theta - b(\theta)}{a(\phi)} + c(\gamma,\phi)\right\} \quad (2.1)$$

- θ é a forma canónica do parâmetro escalar de localização;
- ϕ é um parâmetro escalar de dispersão conhecido;
- $a(.)$, $b(.)$ e $c(.,.)$ são funções reais conhecidas.

O valor médio e a variância da distribuição exponencial obtém-se derivando a função densidade de probabilidade, $\ell(\theta, \phi, y) = \ln(f(y|\theta,\phi))$, designada por função Score:

$$S(\theta) = \partial l(\theta;y,\phi)\ \partial\theta = Y - b'(\theta)\ a(\phi) \quad (2.2)$$

Sabe-se que para famílias regulares se tem:

$$E\big(s(\theta)\big) = 0\ e\ E[s^2] = E\left[\left(\frac{\partial l(\theta;\gamma,\phi)}{\partial\theta}\right)^2\right] = -E\left[\frac{\partial^2 l(\theta;\gamma,\phi)}{\partial\theta^2}\right] \quad (2.3)$$

Assim,

$$S(\theta) = \frac{Y - b'(\theta)}{a(\phi)}\ e\ \frac{\partial S(\theta)}{\theta} = -\frac{b''(\theta)}{a(\phi)}, onde\ b'(\theta) = \frac{\partial b(\theta)}{b(\theta)}\ e\ b''(\theta) = \frac{\partial^2 b(\theta)}{\partial\theta^2} \quad (2.4)$$

E, portanto,

$$E(Y) = \mu = b'(\theta)\ e, Var(Y) = a(\phi)b''(\theta) \quad (2.5)$$

Alguns exemplos de distribuições conhecidas que pertencem à família da exponencial são a Normal, Poisson, Gama e Binomial.

Estimação dos parâmetros[7]

[6] De acordo com Turkman, M. Antónia e Silva, Giovani, Modelos Lineares Generalizados.
[7] De acordo com Turkman, M. Antónia e Silva, Giovani, Modelos Lineares Generalizados.

A inferência com MLG é essencialmente baseada na verosimilhança. Isto é, apesar do método da máxima verosimilhança ser o método de eleição para estimar os parâmetros de regressão, os testes de hipóteses sobre os parâmetros do modelo e de qualidade de ajustamento são métodos de estimação alternativos baseados também na verosimilhança. Considerando uma amostra aleatória $(y_1, \dots, y_n)$ a função log-verosimilhança é dada por:

$$\ell(\theta, \phi; y_1, \dots, y_n) = \sum_{i=1}^{n} \ln f(y_i; \theta, \phi) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right] \tag{2.6}$$

Por fim, partindo do princípio que existe solução e que essa é única, os estimadores de máxima verosimilhança são determinados através da solução das equações de verosimilhança apresentadas:

$$\frac{\partial l(\beta)}{\partial \beta_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{\partial l(\beta_j, \gamma_i)}{\partial \beta_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \left[\frac{\gamma_i - b'(\theta_i)}{a(\phi)} \right] \frac{\partial \theta_i}{\partial \beta_j} = 0, j = 1, \dots, p \tag{2.7}$$

Desvio ou análise do erro

O modelo saturado é útil para julgar da qualidade de ajustamento de um modelo em análise. A distância dos valores ajustados do modelo saturado e dos correspondentes valores observados y é uma medida de discrepância entre o modelo saturado e o modelo corrente. Se compararmos o modelo em análise com o modelo saturado através da estatística de razão de verosimilhanças, obtemos

$$D = 2\phi(\tilde{l} - \hat{l}) \tag{2.8}$$

Onde $\tilde{l}$ é log-verosimilhança do modelo saturado e $\hat{l}$ log-verosimilhança é log-verosimilhança do modelo atual.

Critério de informação akaike e bayesiano

Um dos critérios de comparação de modelos é o designado de Akaike (AIC) (Akaike, 1974) é baseado na função de máxima verosimilhança, definido por:

$$AIC = -2\ell + 2p \tag{2.9}$$

Ao comparar vários modelos, baseados nas mesmas observações, é selecionado o modelo cujo AIC é o menor. O Critério de Informação Bayesiano (BIC), proposto por Schwarz (1978) é dado por:

$$BIC = -2\ell + 2p * \log(n) \tag{2.10}$$

onde p é o número de parâmetros a serem estimados e n é o número de observações da amostra.

Regressão logística

Supondo que dispomos de n observações independentes, $(y_1, \dots, y_n)$ de variáveis aleatórias em que cada uma tem distribuição de Bernoulli de parâmetro pi=p(xi), em que xi, i =1,...,n, são observações de uma variável independente e que vamos tratar como um conjunto de constantes. Assim, a função massa de probabilidade de cada uma das observações é dada por

$$P\{Y_i = y_i | x_i\} = p(x_i)^{y^i}(1 - p(x_i))^{y_i} \quad (2.11)$$

em que cada y_i só pode tomar o valor 0 ou 1.

A distribuição de Bernoulli é um caso particular da família exponencial. O parâmetro natural $logit(\theta)$ = $ln(\frac{\mu}{1-\mu})$, que por sua vez é equivalente a ter $\mu = \frac{e^{\theta}}{1+e^{\theta}}$. Além disso, ϕ=1, $a(\phi) = \phi$, $b(\theta) = -\ln(1 - \mu)$ = $\ln(1 + e^{\theta})$ e $c(y,\phi) = 0$. E portanto, $E[Y] = \frac{e^{\theta}}{1+e^{\theta}} = \mu$ e $Var[Y] = \frac{e^{\theta}}{1+e^{\theta}} = \mu(1 - \mu)$, respetivamente.

Assim sendo, considerando a distribuição de Bernoulli, a variável binária pode ser modelada por um GLM com probabilidade de sucesso igual a μ. Através da função logit temos a seguinte equação $g(\mu)$ = $\theta = \ln(\frac{\mu}{1-\mu})$,). A função logit será usada na regressão logística.

Teste da razão de verosimilhanças

O teste da razão de Verosimilhanças compara os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem a variável em questão. Com o objetivo de garantir que a variável independente é significativa, compara-se o valor de D com e sem a variável na equação. A alteração em D devido à inclusão da variável no modelo é dada por:

$$G = D(modelo\ sem\ a\ variável) - D(modelo\ com\ a\ variável) \quad (2.12)$$

Logo, pode-se escrever a estatística G como:

$$G = -2 \ln\left[\frac{(verosimilhança\ do\ modelo\ sem\ a\ variável)}{(verosimilhança\ do\ modelo\ com\ a\ variável)}\right] \quad (2.13)$$

Teste de Wald

O teste de Wald compara a estimativa de máxima verossimilhança do parâmetro β^j e a estimativa do seu erro padrão. O teste de Wald, realiza-se para as seguintes hipóteses:

$H0$: $\beta j = 0$ vs $H1$: $\beta j \neq 0$, $j = 0, \dots, p$, sendo p as variáveis as testar.

A estatística do teste Wald para a regressão logística é dada por: $Wj = \frac{{\beta_j}^2}{var(\beta_j)}$. A hipótese nula tem distribuição Qui-quadrado com um grau de liberdade.

Qualidade do ajustamento

A matriz de confusão é uma tabela que permite uma visualização simplificada do número de classificações corretas e do número de classificações preditas de uma classe (Han e Kamber, 2006). Uma forma de avaliar a qualidade do ajustamento do modelo consiste em comparar os valores previstos com os valores observados. Quanto maior for o nível de concordância entre os valores observados e previstos, melhor será o modelo. Os valores positivos que são previstos como tal, chamam-se verdadeiros positivos (VP) e os que são previstos de forma errada como negativos são denominados como falsos negativos (FN). Analogamente, os verdadeiros valores negativos podem ser previstos corretamente como sendo negativos, isto é, os verdadeiros negativos (VN), ou podem ser previstos como positivos, os falsos positivos (FP). Após a classificação de todos os pares de valores os resultados da contagem do total de pares nas classes VP, FN, VN e FP são apresentados esquematicamente na designada Matriz de Confusão.

Tabela 2.1 – Matriz de confusão

		Valores observados		**Total**
		Positivos	**Negativos**	
Valores previstos	**Positivos**	VP	FP	VP+FP
	Negativos	FN	VN	FN+VN
Total		P=VP+FN	N=FP+VN	P+N= Nº Observações

A precisão consiste na proporção de identificações positivas realmente correta, dada por:

$$Precisão = \frac{VP}{VP + FP} \tag{2.18}$$

O modelo ideal é aquele cujo número de falsos positivos (FP) e de falsos negativos (FN) é o menor possível. É possível alterar esses valores alterando o ponto de corte, no entanto, não se pode diminuir o número e falsos positivos sem aumentar os falsos negativos, e vice-versa. Assim sendo deve-se tentar manter um equilíbrio, ou apreciar a qualidade do modelo segundo duas probabilidades, a especificidade e a sensibilidade.
A especificidade é a probabilidade de uma observação ser classificada como negativa dado que é efetivamente negativa e é dada por:

$$Especificidade = \frac{VN}{VN + FP} \tag{2.19}$$

A sensibilidade é a probabilidade de uma observação ser classificada como positiva dado que é efetivamente positiva e é dada por:

$$Sensibilidade = \frac{VP}{VP + FN} \tag{2.20}$$

Curva ROC[8]

A curva ROC (*Receiver Operating Characteristic*) é um método gráfico robusto, que pela sua simplicidade permite estudar a variação da sensibilidade e especificidade, para diferentes valores de corte. Esta é representada graficamente pelo complementar da especificidade (1-E) nas abcissas e pela sensibilidade (S) nas ordenadas e, sendo estas medidas de probabilidade, variam entre 0 e 1. O modelo é considerado perfeito quando a sensibilidade e a especificidade são iguais a 1.

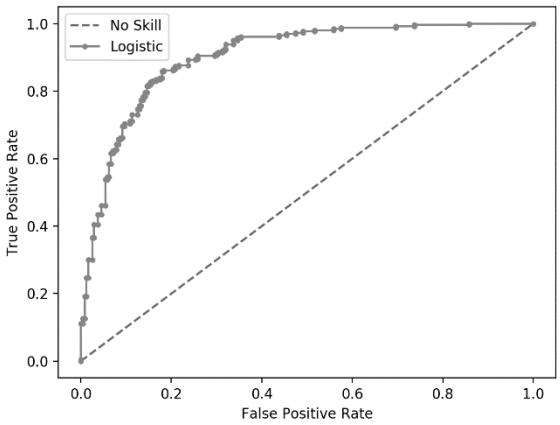

Figura 2.2 – Curva ROC

A curva ROC é comummente utilizada na comparação de modelos. Uma vez que a exatidão de um teste diagnóstico é proporcional à área abaixo da curva (AUC), quanto maior for a área abaixo da curva, maior será a exatidão do modelo. A área abaixo da curva ROC está associada ao poder discriminante de um teste de diagnóstico de um modelo individual e é um dos índices de precisão usados para avaliar a qualidade da curva. Os seus valores estão compreendidos entre 0 e 1. Os modelos mais realistas têm uma AUC superior a 0,5. No entanto, os valores da AUC podem tomar valores dentro de 5 níveis de discriminação:

Tabela 2.2 – Classificação dos valores de AUC

Valor	Classificação
[0,9;1]	Discriminação excecional
[0,8;0,9]	Discriminação excelente
[0,7;0,8]	Discriminação aceitável
[0,6;0,7]	Discriminação fraca
[0,5;0,6]	Não há discriminação

[8] Braga A.C, Curvas ROC: Aspectos funcionais e aplicações, 2000.

Métodos de seleção de variáveis[9]

O método de Backward inicia-se com todas as variáveis do modelo e elimina sucessivamente as variáveis com *p-value* superior ao nível de significância considerado. Por sua vez, o método de Foward inicia-se sem nenhuma variável no modelo e acresecenta sucessivamente as variáveis com p-value inferior ao nível de significância. O método de regressão Stepwise é um método iterativo que resulta da combinação dos métodos de seleção Bacward e Foward e que permite a construção de um modelo estatístico através da inclusão ou exclusão de variáveis, usando testes F ou T. O método Stepwise requer dois níveis de significância: um para adicionar variáveis, o qual se designa o alfa de entrada e outro para remover variáveis, o qual se designa alfa de saída. O alfa de entrada e saída nunca devem ser iguais. O método inicia-se com uma única variável, aquela que tiver maior correlação com a variável dependente. A cada passo do método Foward, após incluir a variável aplica-se o método Backward. Reajusta-se o modelo e vai-se repetindo o processo até que todas as variáveis sejam significativas. O método termina quando todas as variáveis entraram no modelo ou quando todas candidatas a sair no modelo tem um p-valor menor que o alfa de saída e as variáveis não incluídas tem um p-value superior ao alfa de entrada.

[9] De acordo com Alpuim, T, Modelos Lineares- Notas de apoio à disciplina, 2015.

CAPÍTULO III – Contratos continuados, Construção de um Modelo de Renovação

Quando ocorrem sinistros inesperados como tempestades, incêndios em grande escala ou inundações, as companhias de seguros podem não ter valores para acomodar os gastos desses fenómenos extremos. Por outro lado, a rentabilidade pode não ser a esperada, existindo receitas baixas que podem ou não ser lucrativas. Para fazer face a essas adversidades existem processos de *repricing* que consistem no aumento do prémio da apólice consoante as suas caraterísticas. Através da análise dessas caraterísticas verifica-se quais são as que poderão fazer variar o preço. Por exemplo, se uma empresa tiver registado vários sinistros nos últimos anos é normal que o prémio na renovação dessa apólice aumente.

O modelo de Renovação é um modelo aplicado a apólices já existentes na companhia. A variável dependente é a probabilidade de o cliente renovar ou não a apólice, uma variável binária que toma o valor 1 se o cliente renovou o contrato e 0 se não renovou.

Preparação da base de dados

A fase inicial do modelo trata-se de construir a base de dados para a construção do modelo. O objetivo é selecionar previamente as variáveis que se quer testar se são ou não relevantes para explicar a variável dependente. Para isso, teve-se como ponto de partida variáveis de diversas fontes, com base no estudo de modelos de renovação já construídos anteriormente para outros ramos e que parecem ser transversais. Uma vez que a maioria dos dados das seguradoras são obtidos pelo preenchimento do contrato de seguro por parte do cliente, existem erros e por vezes, falta de informação. Por esse motivo, após a construção da base de dados é necessário fazer o tratamento da mesma, analisar as variáveis incluídas, observar a sua distribuição, a existência de possíveis *outliers* ou células vazias. Para a construção deste modelo foram utilizados os dados das renovações de novembro de 2018 a dezembro de 2019.

Seleção de variáveis

Em seguida, apresentam-se algumas das variáveis usadas para estimar a probabilidade de renovação.

Tabela 4.1 – Variáveis em estudo

Variável	Descrição	Tipificação
Variáveis associadas ao cliente		
Número de apólices	Número de apólices que o cliente tem na companhia	Quantitativa
Número de apólices canceladas	Número de apólices canceladas que o cliente tem na companhia	Quantitativa
Número de ramos	Número de ramos que o cliente tem na companhia	Quantitativa
Variáveis associadas à apólice		
Antiguidade da apólice	Número de anos da apólice na companhia	Quantitativa
Número de sinistros	Número de sinistros associados à apólice	Quantitativa
Número de sinistros recusados	Número de sinistros recusados associados à apólice	Quantitativa
Custo do sinistro	Custo do último sinistro associado à apólice	Quantitativa
Fracionamento	Frequência de pagamento	Categórica
Desconto/Agravamento	Valor do desconto ou agravamento aplicado à apólice na anuidade anterior	Quantitativa
Canal comercial	Classificação do mediador que realizou o contrato	Categórica
Rede comercial	Classificação da rede que realizou o contrato	Categórica
Capital seguro	Valor monetário do total do capital seguro associado à apólice	Quantitativa
Prémio comercial	Prestação paga pelo segurado, para a contratação do seguro	Quantitativa
Variáveis associadas ao agente		
Certificação	Indica se o agente responsável pela realização do contrato dispunha de certificação	Categórica
Variáveis demográficas		
Distrito	Distrito da morada do tomador do seguro	Categórica

Teste Qui-quadrado, significância variáveis categóricas

Tabela 4.2 – Teste do Qui-quadrado às variáveis categóricas

Variável	Qui-Quadrado	Graus de liberdade	P-value
Número de apólices canceladas	1017,4	1	<0,0001
Canal comercial	100,3	7	<0,0001
Rede comercial	54,8	6	<0,0001
Número de sinistros recusados	51,5	1	<0,0001
Distrito	48,9	13	<0,0001
Fracionamento	32,6	2	<0,0001
Custo do sinistro	24,0	2	<0,0001
Certificação	19,1	1	<0,0001
Número de sinistros	9,0	1	0,0028
Fator A	1,9	3	0,5872

Numa primeira análise observou-se os valores dos *p-values* resultantes da execução de um teste do Qui-Quadrado, de modo a confirmar se as variáveis categóricas trazidas para o modelo eram ou não relevantes para explicar a variável dependente, a probabilidade do cliente renovar ou não o contrato. Tendo como referência um nível de significância de 5%, a única variável não significativa encontrada foi a variável identificada como fator A, pelo que foi removida. A tabela anterior apresenta uma sugestão inicial da maior ou menor importância de cada variável para explicar a variável dependente.

Correlação

O passo seguinte foi a identificação da possível dependência entre duas variáveis. Para o modelo estar perfeitamente adequado, as variáveis altamente correlacionadas foram removidas do modelo, isto é, as variáveis cujo valor do *V-Cramer* era superior a 0,7.

Tabela 4.3 – Matriz de correlações V-Cramer apenas com as variáveis correlacionadas

Variável I	Variável II	V-Cramer
Prémio comercial	Capital seguro	0,873

Pela matriz de correlações identificou-se a existência de uma dependência entre duas variáveis. Essas foram o prémio comercial e o capital seguro. A escolha da variável a manter no modelo foi feita tendo-se realizado à priori uma categorização destas variáveis quantitativas, seguindo-se uma análise da significância das mesmas através da execução de um teste do Qui-Quadrado. Verificou-se que o capital seguro era a variável mais significativa e, portanto, foi essa que permaneceu no modelo.

Análise Descritiva

Após excluir as variáveis não significativas e variáveis correlacionadas, analisou-se o comportamento da taxa de renovação de acordo com cada variável dependente. É importante reforçar a ideia de que essa análise demonstra a influência de cada variável sozinha com a variável dependente.

Variáveis categóricas

- Fracionamento

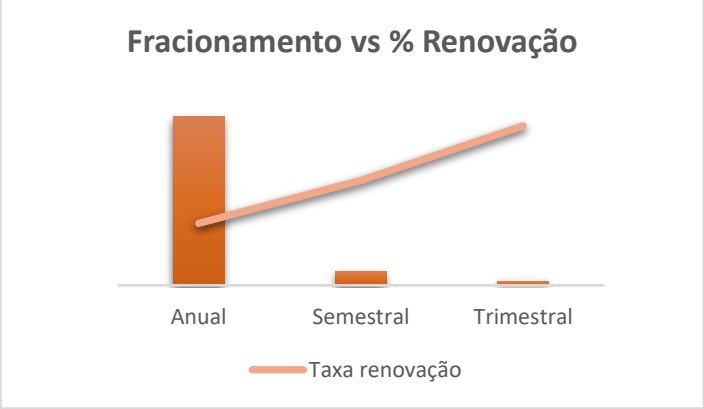

Gráfico 4.2 – Taxa de renovação por fracionamento

O gráfico anterior sugere que os clientes que pagam o seu prémio de forma anual renovam muito menos do que aqueles que pagam semestral ou trimestralmente. Ou seja, a taxa de renovação parece diminuir

quanto maior é a periodicidade entre dois pagamentos. A razão aparente desta tendência é que os clientes que pagam anualmente sentem mais o aumento do que aqueles que pagam semestralmente e trimestralmente. A maioria da carteira Ageas Seguros diz respeito a clientes que pagam anualmente.

- Canal comercial

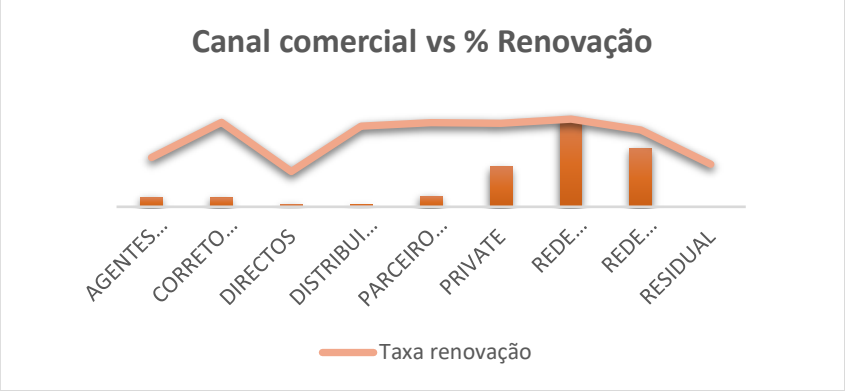

Gráfico 4.3 – Taxa de renovação por canal comercial

Os canais comerciais com mais vendas são os agentes exclusivos, multimarca e *private*. A taxa de renovação mais elevada pertencente aos agentes exclusivos, seguindo-se os corretores e parceiros ativos remotamente. A elevada percentagem nos agentes multimarca deve-se ao facto destes agentes venderem apenas uma marca e, portanto, serem mais persistentes com o cliente para esse adquirir o seguro na companhia à qual pertencem.

- Rede comercial

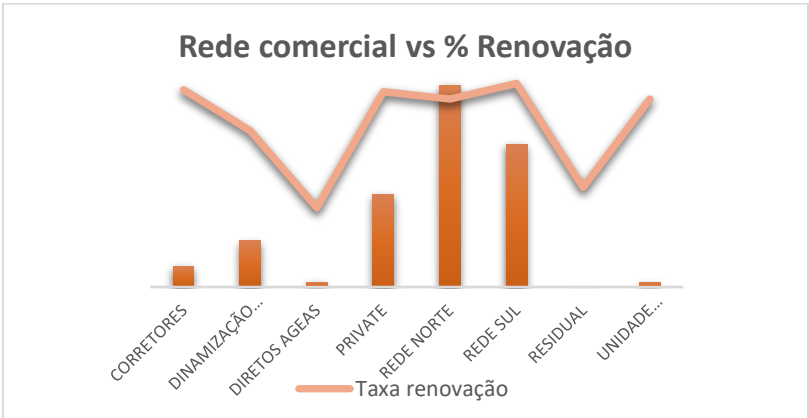

Gráfico 4.4 – Taxa de renovação por rede comercial

O gráfico mostra que as principais redes de venda são a rede norte e a rede sul, verificando-se nessas redes também uns dos maiores valores da taxa de renovação.

- Certificação

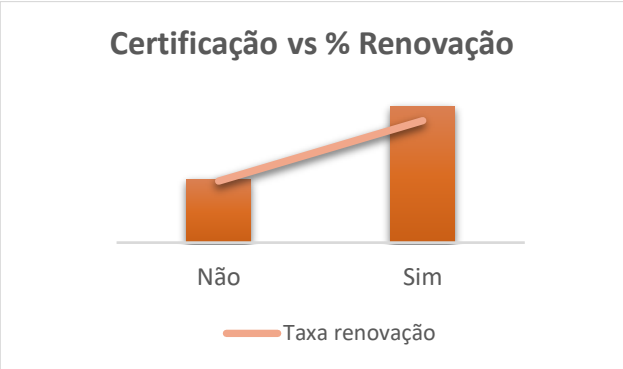

Gráfico 4.5 – Taxa de renovação por certificação do agente

A certificação diz respeito a um teste realizado pela companhia para apostar na formação dos seus agentes. Tal como se esperava, os contratos realizados por agentes com certificação, aparentam ter uma taxa de renovação superior.

- Distrito

Para esse efeito, para a variável distrito, agregaram-se distritos por distância, proximidade entre taxas de renovação e menor exposição.

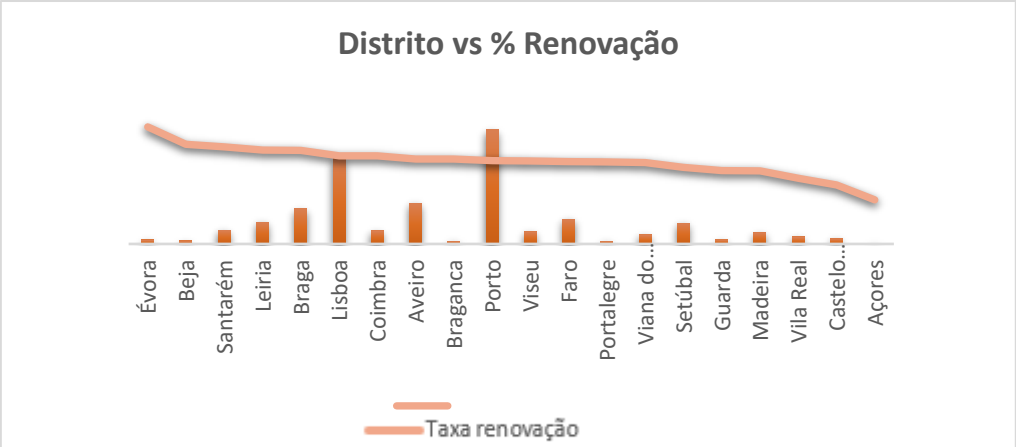

Gráfico 4.13 – Taxa de renovação por distrito

Tabela 4.4 – Agregação dos distritos por taxa de renovação

Distrito	Taxa de renovação	Variação da taxa de renovação
Évora	94,9%	
Beja	91,9%	3,2%
Santarém	91,5%	0,5%
Leiria	91,0%	0,6%
Braga	90,9%	0,1%
Lisboa	90,0%	1,0%
Coimbra	90,0%	0,0%
Aveiro	89,5%	0,6%
Bragança	89,4%	0,1%
Porto	89,2%	0,3%

Viseu	89,1%	0,0%
Faro	89,0%	0,1%
Portalegre	88,9%	0,1%
Viana do Castelo	88,8%	0,1%
Setúbal	88,1%	0,9%
Guarda	87,5%	0,6%
Madeira	87,4%	0,1%
Vila Real	86,2%	1,4%
Castelo Branco	85,0%	1,4%
Açores	82,5%	3,0%

Tabela 4.5 – Distritos agregados

Distrito	Taxa de renovação
Beja e Évora	93,5%
Santarém e Leiria	91,2%
Braga	90,9%
Lisboa	90,0%
Aveiro e Coimbra	89,6%
Viseu e Bragança	89,2%
Porto	89,2%
Faro	89,0%
Portalegre e Viana do Castelo	88,8%
Setúbal	88,0%
Guarda e Madeira	87,4%
Vila Real	86,2%
Castelo Branco	85,0%

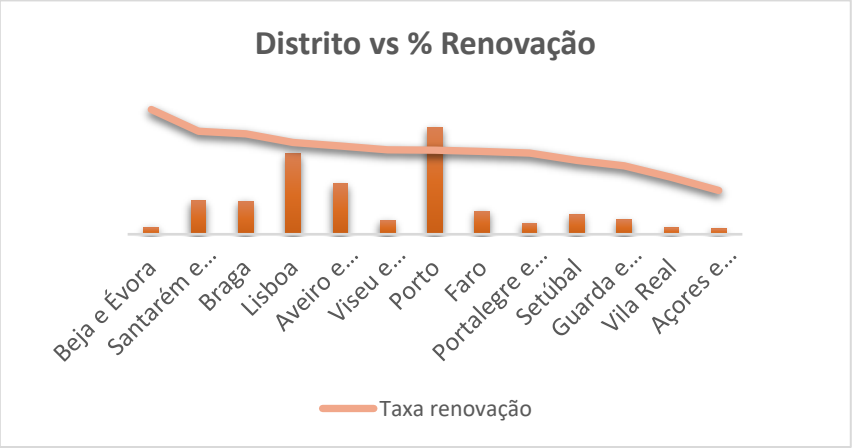

Gráfico 4.14 – Taxa de renovação por distritos agregados

Os agrupamentos de distritos que mais convertem são as grandes metrópoles Lisboa e Porto, seguindo-se Aveiro e Coimbra. Por outro lado, os Açores e Castelo Branco são dos distritos com menor exposição e também menor propensão à renovação.

Variáveis quantitativas

- Número de apólices

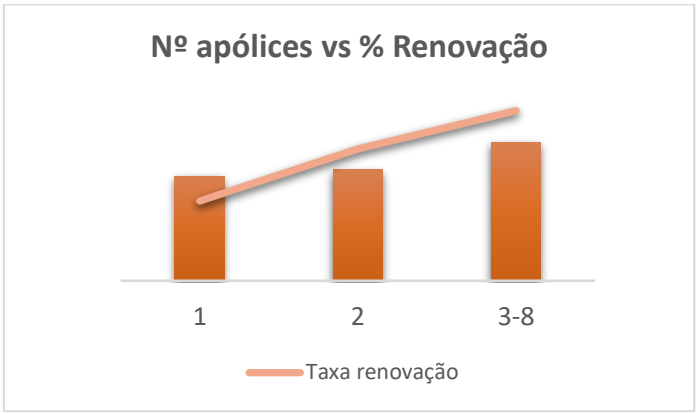

Gráfico 4.17 – Taxa de renovação por nº de apólices

Aparentemente, quando o número de apólices aumenta, a tendência da taxa de renovação é aumentar também. Quando o cliente possui mais de uma apólice na companhia, a sua vinculação à mesma é maior e, portanto, este opta por renovar o contrato.

- Número de apólices canceladas

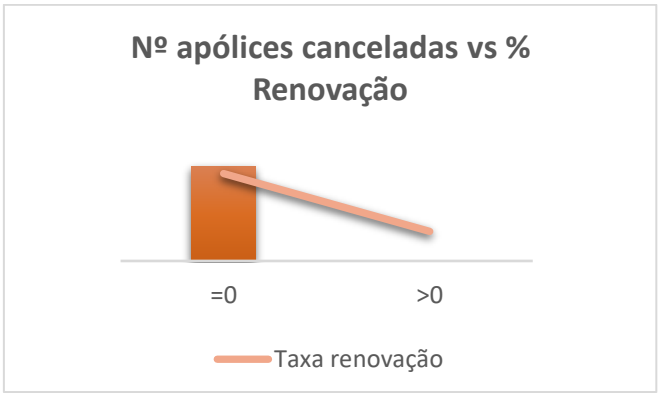

Gráfico 4.18– Taxa de renovação por nº de apólices canceladas

O número de apólices canceladas foi agrupado em duas categorias, igual a 0 e maior que 0, contudo, como podemos ver no gráfico acima, a maioria dos clientes possui 0 apólices canceladas. A taxa de renovação é muito menor em clientes com mais de 0 políticas canceladas.

- Número de ramos

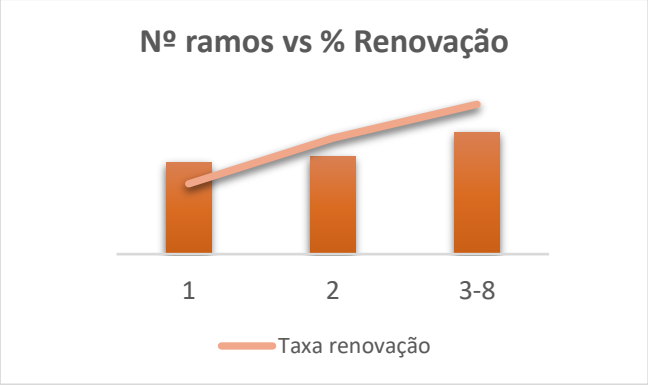

Gráfico 4.19– Taxa de renovação por nº de ramos

Quando o número de ramos aumenta, a tendência da taxa de renovação é aumentar. Quanto mais ramos o cliente tiver, maior é a sua confiança na empresa e maior é a probabilidade de renovação.

- Antiguidade da apólice

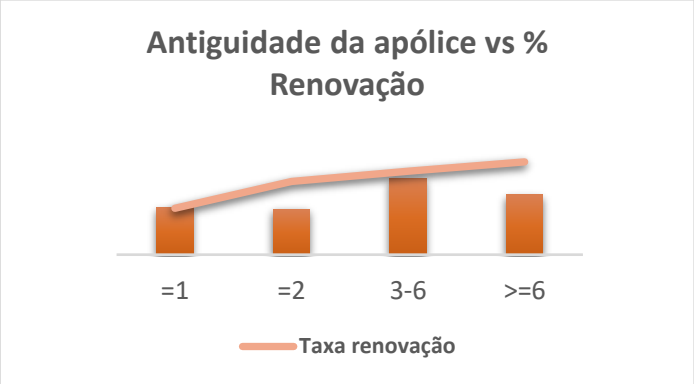

Gráfico 4.20 – Taxa de renovação por antiguidade da apólice

O gráfico da taxa de renovação indica que a tendência da taxa de renovação é aumentar conforme o número de anos da apólice aumenta. Ou seja, se o cliente for renovando consecutivamente o contrato significa que a sua confiança na companhia está a aumentar e, portanto, a probabilidade de renovar na anuidade seguinte é maior.

- Número de sinistros

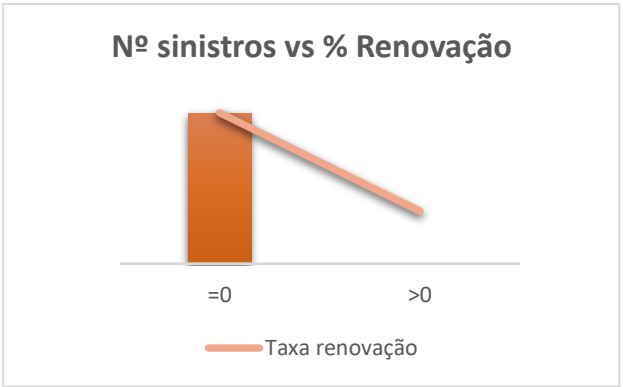

Gráfico 4.22– Taxa de renovação por nº de sinistros

Decidiu-se agrupar a variável em duas categorias, igual a zero ou maior que zero. Como era esperado, os clientes sem sinistros renovam mais.

- Número de sinistros recusados

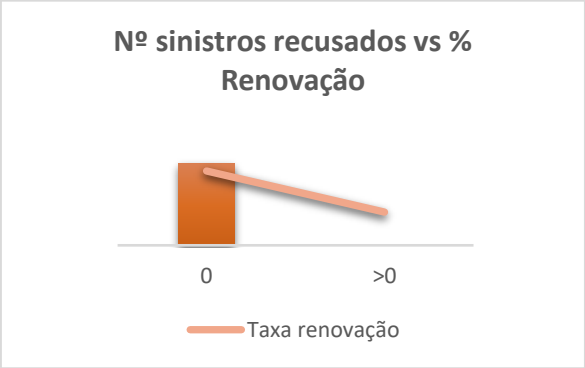

Gráfico 4.23– Taxa de renovação por nº de sinistros recusados

Decidiu-se agrupar a variável em duas categorias, igual a zero ou maior que zero. Como era esperado, os clientes sem sinistros recusados renovam mais.

- Custo do sinistro

Gráfico 4.24 – Taxa de renovação por custo do sinistro

O custo dos sinistros foi previamente agrupado em três categorias, igual a 0, entre 0 e 1000 e maior que 1000. Essa divisão foi feita com base no conhecimento de negócio e, como é possível confirmar, os clientes comportam-se de forma diferente nos três grupos. Há um declínio da taxa de renovação no custo de sinistros entre 0 e 1000 euros. O que o gráfico aparenta é que os clientes com baixo valor do custo de sinistros não se sentem tão satisfeitos com o seu atendimento.

- Desconto/Agravamento

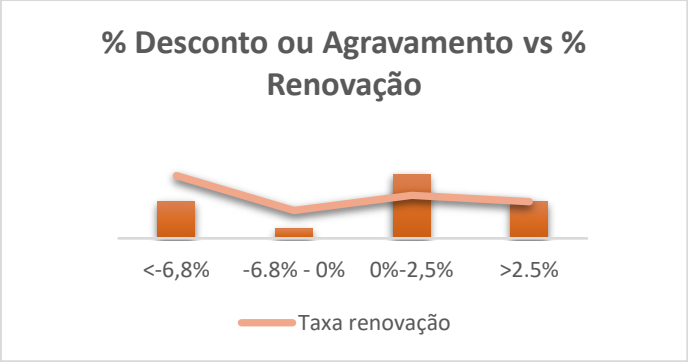

Gráfico 4.25 – Taxa de renovação por % de desconto ou agravamento

Quando o cliente tem um pequeno desconto, não o parece sentir, uma vez que a taxa de renovação é menor do que quando existe um pequeno aumento. Quando o desconto é maior, a taxa de renovação aumenta consideravelmente.

- Capital seguro

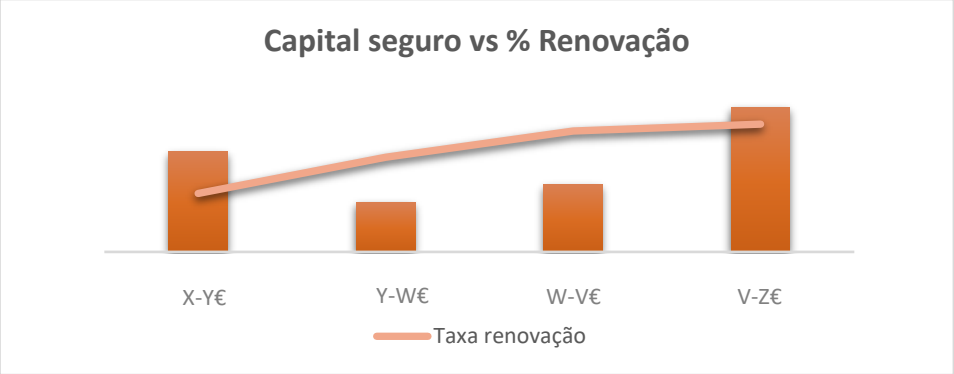

Gráfico 4.26 – Taxa de renovação por valor de capital seguro

O capital da apólice refere-se ao valor do seguro do imóvel, ao conteúdo ou a ambos. Analisando o gráfico, podemos ver que os clientes com um capital seguro mais alto (Z>V>W>Y>X) têm maior probabilidade de renovação. Uma vez mais, o vínculo à empresa assegura na maioria das vezes uma maior propensão à renovação.

Modelação

Árvores de decisão

Na fase de modelação, criou-se uma árvore de decisão simples com o objetivo de identificar as caraterísticas que mais influenciam a renovação através de subdivisões. Esta é uma abordagem comportamental que sustenta os modelos obtidos pela regressão logística.

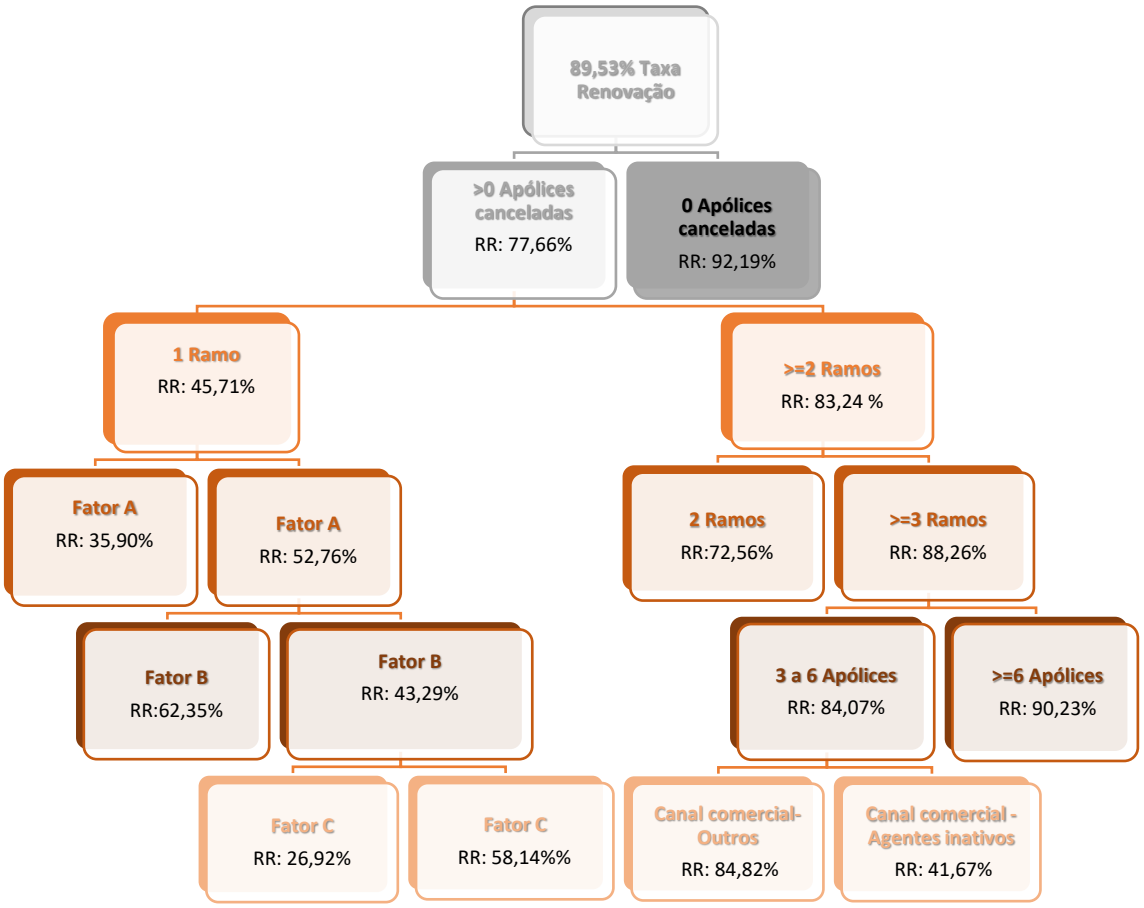

Figura 4.1 – Árvore de decisão resultante

A árvore de decisão obtida assinala que clientes com 0 apólices canceladas são o perfil com a maior taxa de renovação. Além disso, observa-se que para clientes com mais de 0 apólices canceladas, 3 ou mais ramos e mais de 6 apólices têm uma maior probabilidade de renovação. Isso significa que os clientes com um relacionamento maior com a empresa, isto é, mais riscos em vigor e mais apólices, acabam por ser os mais propensos a renovar.

Regressão Logística

Utilizaram-se três métodos de seleção de variáveis para a construção do modelo final. Esses foram o método Backward, o método Foward e o método Stepwise, sendo que o critério de escolha foi selecionar o método do que resultasse o modelo com menor AIC. O valor de significância utilizado foi de 5% (α=0.05).

Após aplicar os três métodos concluiu-se que todos resultaram no mesmo modelo final, isto é, com as mesmas variáveis significativas e com um AIC associado de 12.022,721.

Tabela 4.8 – Desenvolvimento do método Stepwise

Iteração	Parâmetro	Graus de liberdade	Qui-Quadrado	P-value
1	Número de apólices canceladas	1	715.0311	<0.0001
2	Número de apólices	3	746.9463	<0.0001
3	Fator A	2	134.8030	<0.0001
4	Capital Seguro	3	98.3569	<0.0001
5	Número de ramos	2	83.3281	<0.0001
6	Antiguidade da apólice	3	48.8482	<0.0001
7	Número de sinistros recusados	1	36.6827	<0.0001
8	Fator B	11	51.9860	<0.0001
9	Canal commercial	7	33.5990	<0.0001
10	Fator C	3	21.4889	<0.0001
11	Fracionamento	2	12.8575	0.00016
12	Fator D	3	11.8372	0.0080
13	Distrito	12	25.2580	0.0136
14	Fator E	4	11.6162	0.0204

A tabela acima apresenta o valor das estatísticas de teste, a ordem pela qual as variáveis entraram no modelo logístico e o seu *p-value*. Segundo o método Stepwise o número de apólices canceladas é a variável mais significativa, conclusão sustentada pelos resultados obtidos pela árvore de Decisão.

Seja PR a probabilidade de renovação, o modelo final é definido pela seguinte expressão:

$$\text{Logit(PR)} = \beta_0 + \beta_1 * \text{variável\_1} + \beta_2 * \text{variável\_2} + \ldots + \beta_{58} * \text{variável\_58}$$

Em que os valores de β são, uma vez mais, confidenciais.

Validação do modelo

O próximo passo foi confirmar se os novos dados se encaixavam e avaliar a qualidade do ajustamento. Quando a base de dados foi criada, foi dividida aleatoriamente entre a base para modelação (70%) e a base para teste (30%). Os dados de teste foram usados para entender o quão bem o modelo selecionado se ajustou quando aplicado a novos dados. Para esse efeito analisou-se a curva ROC e a matriz de confusão.

- Curva ROC

Para concluir sobre a capacidade discriminatória do modelo, calculou-se a área abaixo da curva ROC, denominada AUC (*Area Under The Curve*). Quanto maior o valor da AUC, melhor o modelo explica os dados.

Em seguida apresentam-se as curvas ROC para a base de treino e para a base de teste. O AUC da base de treino teve um valor de 0.756 enquanto o AUC da base teste teve um valor de 0.737, logo, ligeiramente inferior. Assim sendo, de acordo com a tabela de classificação de Hosmer-Lemeshow, podemos afirmar que o modelo tem uma capacidade discriminatória aceitável.

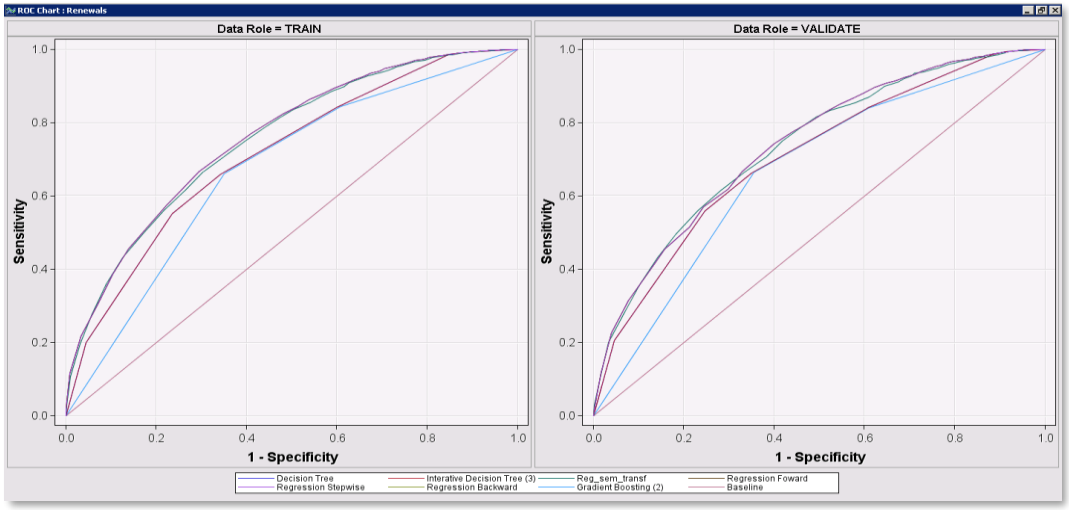

Figura 4.2 – Curvas ROC para a base de treino e de teste, respetivamente

Tabela 4.9 – Matriz de confusão para a base de treino

		Valor observado	
		Positivo	**Negativo**
Valor previsto	**Positivo**	18332	1980
	Negativo	172	96

Tabela 4.10 – Matriz de confusão para a base de teste

		Valor observado	
		Positivo	**Negativo**
Valor previsto	**Positivo**	7857	858
	Negativo	65	42

Através da matriz de confusão é possível calcular a precisão do modelo. Esta é a métrica de avaliação que nos permite verificar dentre todas as classificações positivas que o modelo fez, quantas estão corretas. Se a precisão for alta, pode-se dizer que o modelo está bem classificado.

A base de treino demonstrou ter uma precisão de 90,3% enquanto que a base de teste uma precisão de 90,2%, logo apresentam valores muito próximos, o que indica que o modelo tem uma elevada fração de casos previstos corretos.

Perfis de renovação

Para identificar os perfis de renovação, usou-se um novo conjunto de dados, utilizando para esse efeito os meses de janeiro e fevereiro de 2020. O modelo previu a taxa de renovação para o novo conjunto de dados e identificou 7 perfis de renovação diferentes. O 1º perfil é aquele com maior taxa de renovação, seguido pelo 2º e assim por diante. Logicamente, o 7º perfil é aquele com a menor taxa de renovação. Os perfis não são completamente disjuntos, pois resultam da combinação de várias variáveis. Portanto, apenas se pode obter uma ideia global da sua constituição analisando a seguinte tabela. A taxa geral de renovação observada para esse conjunto de dados foi de 93,9% e o modelo escolhido previu uma taxa geral de renovação de 90,1%. Isso significa que há uma diferença de 4% entre os valores observados e os previstos

Tabela 4.11 – Perfis de renovação

	1º Perfil		2º Perfil		3º Perfil		4º Perfil		5º Perfil		6º Perfil		7º Perfil	
	Pesos		Pesos		Pesos		Pesos		Pesos		Pesos		Pesos	
Nº apólices canceladas	R	NR	R	NR	R	NR	R	NR	R	NR	R	NR	R	NR
=0	89%													
>0			11%			49%		32%		2%		1%		16%
Nº apólices														
1	19%													16%
2	20%							32%						
3-6	31%		3%			30%				2%		0,4%		
>=6	19%		8%			19%				1%		0,4%		
Capital Seguro														
<X €	23%		1%			14%		6%		0,4%				3%
X-Y €	21%		2%			10%		12%		1%		0,4%		6%
Y-W €	24%		3%			12%		8%		0,4%		0,4%		4%
>= W €	22%		4%			14%		7%		0,4%				3%
Nº ramos														
1	27%							8%		2%				16%
2	29%					49%		24%				1%		
>=3	33%		11%											

Com o intuito de identificar cada um dos perfis de renovação escolherem-se as cinco variáveis que o modelo final selecionou como as mais significativas: o número de apólices canceladas, o número de apólices, a forma de pagamento, o capital seguro e o número de ramos. Ao analisar a tabela é importante observar que o 1º e o 2º perfil têm a maior % de renovações, e os perfis 3, 4, 5, 6 e 7 têm a maior % de não renovações.

Em relação ao número de apólices canceladas, o fato de o 1º perfil ter a maior taxa de renovação e estar associado a 0 apólices canceladas, indica que os clientes com maior probabilidade de renovação tendem a ter 0 apólices canceladas.

Quando se analisa o número de apólices, os perfis que renovam menos estão mais associados aos clientes com 1 ou 2 apólices e os clientes com 3 a 6 apólices têm maior probabilidade de renovação. Isso reforça a ideia de que clientes com maior relacionamento com a empresa tendem a renovar mais.

No que diz respeito ao capital seguro, os clientes com um capital seguro pertencente às diferentes classes têm um comportamento de renovação bastante semelhante. Em termos de comportamento de não renovação, observa-se que os clientes que não renovam têm frequentemente um valor segurado entre X € e Y € (W>Y>X).

Em relação ao número de ramos, os clientes que com três ou mais ramos tendem a renovar mais. Os clientes com um ou dois ramos tendem a renovar menos, como se pode ver no 3º e 7º perfil, pois são perfis que renovam menos. A percentagem de não renovação é de 49% e 16%, ou seja, altos níveis de taxa de não renovação.

Capítulo IV - Conclusão

Foi uma honra poder ter iniciado a minha primeira experiência profissional a realizar este projeto numa empresa internacional de renome e num mercado que também tive a oportunidade de conhecer na realização do meu mestrado, o mercado segurador. As sinergias entre o que aprendi em contexto académico e profissional foram imensas e permitiram-me terminar este projeto com uma bagagem de aprendizagem fundamental para o meu futuro. Espero, também eu, ter contribuído para o aumento da rentabilidade da Ageas Seguros.

No âmbito dos contratos continuados utilizou-se a regressão logística para modelar a probabilidade de renovação no ramo em estudo. Avaliou-se a capacidade de o modelo prever corretamente o perfil das apólices propícias a renovar. Este modelo permite calcular a probabilidade de renovação estimada para cada cliente. A fase que exigiu mais tempo no projeto foi a construção da base de dados e é fundamental para garantir uma base bem modelizada e correta. Retiraram-se ideias bastante interessantes da influência de cada variável na taxa de renovação. Concluiu-se que o modelo apresenta uma capacidade de previsão global aceitável, medida pelo AUC e pelo valor de precisão. Após a construção do modelo, foi possível dar recomendações para o futuro à área comercial. Uma das conclusões mais importantes a extrair é que o vínculo do cliente com a companhia é um fator muito importantes para a decisão do cliente renovar ou não a apólice.

Bibliografia

[1] Sobre o mundo Ageas. (outubro, 2019). Disponível em:
https://www.grupoageas.pt/

[2] Glossário Ageas, Glossário de Seguros (outubro, 2019). Disponível em:
https://www.ageas.pt/glossario-de-seguros

[3] ASF (2015) – Guia de Seguros e Fundos de Pensões (3º edição) (novembro, 2019). Disponível em:

https://www.asf.com.pt/NR/rdonlyres/3F64D61A-BCDE-48F3-8C36-244F65CD9CAE/0/GuiadeSeguroseFundosdePens%C3%B5es_2015.pdf

[4] Turkman, M. A. A. & Silva G. L. (2000). Modelos Lineares Generalizados – *Da Teoria à Prática.*

[5] Alpuim, T. (2018). *Notas de Apoio à disciplina de Modelos Lineares*

[6] Gomes, J. (2016). *Apontamentos à disciplina de Estatística Aplicada.*

[7] Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression* (2ª Edition). New York: USA: A Wiley-Interscience Publication, John Wiley & Sons Inc.

[8] Mendenhall, W. & Sinich, T. (2011). A Second Course in Statistics – Regression Analysis (7ª Edition). Pearson.

[9] Mitchell, T. (1997). *Machine-Learning.* McGraw-Hill.

[10] Breiman, L. & Friedman, J. & Stone, C. J. & Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis LTD.

[11] SAS Institute Inc (2011). *Getting Started with SAS® Enterprise Miner™ 7.1*. Cary, NC, USA: SAS Institute Inc.

[13] Braga, A. (2000). *Curvas ROC: Aspectos Funcionais e Aplicações*. Tese de Doutoramento, Universidade do Minho